9-2012

An Empirical Examination of the Associations between Social Tags and Web Queries

Kwan Yi
Eastern Kentucky University, kwan.yi@uky.edu

Chan Yun Yoo
University of Kentucky, chan.yoo@uky.edu

informationresearch

VOL. 17 NO. 3, SEPTEMBER 2012

An empirical examination of the associations between social tags and Web queries

Kwan Yi
Department of Curriculum and Instruction, Eastern Kentucky University, Richmond, Kentucky, USA
Chan Yun Yoo
School of Journalism and Telecommunications, University of Kentucky, Lexington, Kentucky, USA

Abstract

Introduction. We aim to discover the associations between social tags for a Web page and Web queries that would retrieve the same Webpage in three major search engines.
Method. 4,827 query terms were submitted to the three major search engines to acquire search engine results pages. A series of Perl scripts were written to read search engine results pages and to identify, analyse, and extract organic links
Analysis. Web pages from the organic links in search engine results pages were examined to see whether and how they had been tagged in Delicious. Only the Webpages tagged by at least 100 taggers were included in this study. The top thirty popular social tags used were harvested. The two sets of data were quantitatively analysed to investigate the research questions.
Results. At least 60% of search engines' query terms overlapped with social tags in Delicious; higher ranked social tags were more likely to be used as query terms for the same Web resources; and the co-occurring pattern of query terms and social tags over social ranking resembled a power law distribution.
Conclusions. Socially tagged resources are likely to be highly ranked in search engine results pages. The findings can be applicable to the future study of Web resource related tasks such as Web searching and Web indexing.

CHANGE FONT

Introduction

Web searching and navigation by search engines have become ubiquitous activities for many people (Pew Internet Project 2010). In a Web search, users put queries into search engines. These queries reflect information needs or search goals of users. Thus, understanding Web queries has been a major focus of research in Web information retrieval. Previous studies sought to classify Web queries using search logs (Jansen *et al.* 2000; 2005), examine implicit user motivations or information needs in search queries, through the analysis of interactions with search engine results (Broder 2002; Chang *et al.* 2006; Hu *et al.* 2009; Lee *et al.* 2005; Rose and Levinson 2004), expand query terms for query modification and relevance feedback, (Collins-Thompson and Callan 2005; Cui *et al.* 2003; Xu *et al.* 2009), and suggest relevant queries to users, (Baeza-Yates and Tiberi 2007; Mei *et al.* 2008; Song *et al.* 2011). Nevertheless, understanding and interpreting Web queries for the purpose of enhancing search engine

performance still remains a challenge in Web retrieval, primarily because of the inherent nature of queries, i.e., only a few words are presented in most queries.

With the recent popularity of Web 2.0 applications, a new type of metadata for Web resources, social tags, have been created and are available for research. In tagging Web resources, people annotate the resources and assign terms (i.e., tags) to the resources. As a result, a collection of tags assigned by different users becomes a cloud of tags called social or collaborative tags. A key benefit of social tagging comes from a cloud of tags, not from an individual tag, with the assumption that a social consensus is formed and realized in the cloud when a considerable number of people participate in a same tagging activity. Such a consensus is described as collective intelligence derived from people. a power of participatory collaboration of a crowd of people (O'Reilly 2005). Being aware of this, in this study we attempt to examine whether a social consensus in a cloud of social tags can be used to help understand and interpret Web queries.

Web searching and social tagging are distinct activities, although both might come from people to identify information. Web queries are typed-in to articulate online users' information needs aiming to locate relevant Web resources. Social tags, however, are produced to express personal annotations of Web resources and are generally used for sharing purposes. Nevertheless, Web queries and social tags are common in describing Web resources using natural language - the vocabularies of information users and providers. Therefore, the fundamental assumption of this study is that there is an association between Web queries and social tags for some Web resources. This study focuses on comparing Web queries submitted for, and social tags assigned to, the same Web resources and investigating the similarity and overlap between the two. There are a number of Web applications related to Web resources such as Web indexing, Web query prediction, and search engine optimization. Our study does not directly aim at such Web applications. However, the results of this study provide some implications to the future study of the Web applications.

Background and related work

Social tagging and social tags

Tagging refers to a practice of assigning keywords (i.e., tags) to information resources. Tagging encourages voluntary participation and this allows for social tagging or collaborative tagging. A primary goal of social tagging is organizing and sharing online information resources. Some popular social tagging or bookmarking applications can be found in Delicious for Web resources, Flickr for images, LibraryThing for books, Technorati for blogs, and Connotea for bibliographic data. Social tagging can be viewed as a new approach to organizing Web resources in that the organization mechanism is not pre-designed but self-directed by the crowd of all participants. The underlying principle on the self-governed mechanism is that the aggregation by all people participated in tagging works better than any single individual's contribution. This is the reason why social tagging is described as a democratic process or task harnessing collective intelligence (O'Reilly and Battelle 2009).

Tags, as keywords assigned to information resources, are the outcome of individual or social tagging efforts. Multiple keywords are generally allowed, as tags, for a single Web resource. However, acceptable types and formats of each tag are subject to the rules implemented in specific tagging systems or applications. For example, in Delicious, a tag is described as a one-word descriptor, and a phrase is not allowed as a single tag. In Delicious, a phrase is treated as a number of individual words, each of which would be a tag. On the contrary, in LibraryThing, placing a space between two words is allowed in a tag, which means that a phrase is permitted as a single tag.

A number of studies have analysed social tags from linguistic and functional perspectives. Heckner, Muhlbacher, and Woff (2007) examined 2,000 social tags from Connotea, and reported that 72% of single-word tags were nouns, 15% were acronyms, 12% were adjectives, and 1% were numbers. But, no adverb or verb forms of tags were used. From the functional perspective, they also reported that 94% of the tags were subject-related, and 96% fell in general descriptors about the content of bibliographic works. In the study of Delicious tags, Kipp and Campbell (2006) reported popular spelling-variations and prevalent use of acronyms and synonyms. The most common spelling-variations were found in the instances of capitals vs. lower cases, singulars vs. plurals, and British English vs. American English spellings. Spiteri (2007) analysed various social tags from Delicious, Furl (now diigo), and Technorati, and found the dominant use of single word tags over multiword tags and that of noun tags over other grammatical forms. She also reported the inconsistent tag choices in tagging between singular and plural words, and the frequent selection of abbreviations and acronyms.

Social tags and indexing

The motivations for social tagging are diverse including; altruistic sharing with others, selfish or personal interests, recordkeeping, organizing, and accessing resources (Ding *et al.* 2010; Hammond *et al.* 2005). Moreover, from the indexing point of view, social tagging is the very opposite to the traditional way to organize the data based on controlled vocabularies, i.e., the choice of indexing terms is guided by a set of rules with a pre-defined list of indexing terms. Thus, the value of social tags as

indexing terms has recently received academic attention. Spiteri (2007) examined whether or not social tags conformed to the guidelines of a controlled vocabulary, a list of controlled terms for a specific community and purpose where terminological relationships are explicitly defined. She reported that the examined tags were fairly consistent with the National Information Standards Organization guidelines in the structure of tags. Even so, she suggested the use of formal guidelines or instruction on how to create and disambiguate tags. Yi (2009) investigated the potentials of social tags as indexing terms in the study in which social tags were used to classify tagged resources into Dewey Decimal Classification using latent semantic analysis. Yi (2009) found that the automated classification task based on such analysis excelled in mean average precision, indicating a significant level of agreement between the automated classification and the adopted manual classification. The results further suggested the importance of social tags as indexing terms because social tags were exclusively used as input data for the automated classification in the study.

Meanwhile, Heckner *et al.* (2007) compared social tags and professionally-driven indexing vocabularies, such as controlled vocabulary and bibliographic metadata, and found that only 54% of social tags were in any part of the bibliographic metadata, i.e., title, abstract, author-assigned keywords and full-text. The greatest overlap between social tags and the metadata was found in titles, followed by full text, author-assigned keywords, and abstracts. The average number of user-assigned tags for a bibliographic item was 2.2, whereas the average number of author-assigned keywords for an item was 5.5. They also reported that users tended to choose either more general or specific terms (that is, different levels of specificity) than the authors for corresponding bibliography. Yi and Chan (2009) compared Delicious tags with Library of Congress Subject Headings and found that approximately two-thirds of all social tags were matched with at least one heading in the Headings, with an additional ten percent of the remaining tags having potential matches. Similarly, Carman (2009) examined the extent to which LibraryThing tags matched their equivalent Library of Congress Subject Headings, and found 52% of tags were matched or closely matched.

Social tags and Web search

From the beginning of the World Wide Web, the textual contents of Web pages have been a primary source of indexing Web content. Later, link structure (i.e., structure of hyperlinks connecting Web pages) and anchor text (i.e., clickable text associated with hyperlinks) have been also used to improve Web indexing and Web searching (Craswell *et al.* 2001; Eiron and McCurley 2003; Page *et al.* 1999). Meanwhile, another approach is using metadata that Web page creators have embedded on their own Web pages (Drott 2002; Lawrence and Giles 2000). In this approach, Web page creators add keywords for Web pages using the <meta /> HTML tag. META keywords provide more information about the Web page for search engines so that search engines adopt the keywords as indexing terms for the Web page. Although adding META keywords can be inappropriately used, such as the repetition of same words, search engines encourage Web site owners and creators to use META keywords to tag their Web sites, which will be displayed in search engine results pages (Bassett and Kumaran 2008). Particularly, Google supports this by the name of Google Co-op (Schwarz 2006). Hence, in this approach, the vocabulary of information providers (i.e., Website creators) is used for indexing the Web, which may be different from the vocabulary of information users (i.e., Web searchers). The discrepancy of using terms between information creators and information users may often lead to low precision in search performance (i.e., too many irrelevant returned pages). In that both social tags and Web queries are the vocabulary of information users, the consideration of linking social tags and Web queries appears to be an appealing research task.

Social tagging is a relatively new phenomenon that has produced a new type of metadata about Web resources. There has been research that compares social tags with keywords occurring at web pages. Li *et al.* (2008) found that important words on web pages are generally covered by the social tagging vocabulary. They also reported that most of the mis-matched words are misspelled or invented by users. It was concluded that there was a consistency between social tags and web pages notated by the tags. The value of social data has been examined in Web information retrieval, specifically for Web search enhancement and Web query modification. For example, Heymann *et al.* (2008) examined the levels of overlap between social tags and search results in results pages. They analysed social tags at over 20,000 posts from Delicious and reported that they were present in the page text of 50% of the annotated Web pages and in the titles of 17% of the pages. Furthermore, they found that 19% of top ten search results from AOL and Yahoo! search engines are tagged in Delicious and 9% of the top 100 search results are in Delicious. Biancalana and Micarelli (2009) studied a way of using social tags, collected from Stumbleupon and Delicious, for the expansion of Web queries. The result of the study demonstrated the effectiveness of using social tags for the expansion in a Web search context. Bao *et al.* (2007) proposed two new relevance page-ranking algorithms based on social data, called SocialSimRank and SocialPageRank. They incorporated socially annotated data in calculating Webpage ranking algorithms, and demonstrated significant improvement in Web searching with the proposed algorithms. Yanbe *et al.* (2007) also proposed a way of combining social tags in a page-ranking algorithm and demonstrated that search systems combining social tags had enhanced the Web search and provided advantages in terms of the results. Chen and Zhang (2009) also dealt with improving the ranking of search result and used tagging information in indexing Web documents. The results of this study showed search performance was enhanced at least 10% when Web documents were indexed by social data.

Heymann *et al.* (2008) claimed that the scale of social data had not grown enough to reach the scale of the current Web. It may be still true at this time. However, online social tagging systems have recently attracted more participants and visitors (eBizMAB

2011; Seaver 2007) and they have continued to grow; nonlinear growth of some tagging systems has been reported (Cattuto *et al.* 2009; Wu 2011). It is expected that the continuing growth of social tagging systems will have an increasing impact on enhancing the quality of Web resource-related applications. It is this expectation that motivates this study. As an extension of the previous studies, we directly compare user-generated tags and Web queries for the same results pages, and examine the relation between social data and Web query.

Research questions

With the recent rise of Web 2.0, socially annotated data such as social tags are quickly emerging, and these have been used in academic studies focusing on improving the quality of Web search. Prior studies have explored the overlap between social tags and content of Web pages (Heymann, *et al.* 2008; Li, *et al.* 2008), expanded Web query using social tags (Biancalana and Micarelli 2009), incorporated social tags into new Web ranking algorithms (Bao *et al.* 2007). Collectively, the academic studies indicate that, social tags have become an important information component in Web searching, along with the content of Web pages and Web queries. However, few have investigated relationships or associations between social tags and Web queries, which we were attempting to answer at this empirical study.

Our fundamental assumption is that a certain level of association may exist between social tags and Web queries. This assumption leads to the first research question, which serves the basis of this study. Examining this question is critical to advance our investigation on more detailed associations (or disassociations) between social tags and Web queries. Although both social taggers and Web searchers attempt to understand and interpret the same Web resources, they may have different perspectives and motivations. Thus, the second and third research questions were suggested to investigate how social tags and Web queries are matched, and ranked in relation to each other. Finally, it is of great importance to examine what contributes the disassociations between social tags and Web queries, because the results may offer insights on how to improve Web search quality: therefore, research question four was suggested. Overall, this study focuses on comparing the social tags for a Webpage in Delicious with the Web queries that would retrieve the same Webpage in search engines.

More specifically, the following research questions are suggested:

Research question 1. Cross coverage of Web resources between Delicious.com and search engine results pages: To what extent are the Webpages in results pages also found in Delicious.com?

Research question 2. Overlap between Web queries and social tags: To what extent do the terms in Web queries that retrieve pages in search engine results pages overlap with social tags for the same pages?

Research question 3. Ranking association between Web queries and social tags: How are Web queries that retrieve pages in search engine results pages ranked within the lists of social tags for the same pages?

Research question 4. Discrepancy analysis: Where do the discrepancies between Web queries and social tags come from?

Figure 1: Tagged Web page in Delicious for the Web page of http://www.delicious.com

Research design

Data collection

Web queries. A limited number of real Web search engine transaction log files are available. The most recent Web log file available, AltaVista 2003 Web query log file, was used in this study. The original query log file contains a total of 3,518,498 individual queries. It contains 1,614,823 unique queries including null queries (see Jansen *et al.* (2005) for data descriptions). In this study, we used four query types (i.e., one-word, two-word, three-word, and four-word queries) for analyses (Jansen *et al.* 2000; Jansen and Spink 2006). The log file contains 1,201,402 unique queries of the four types. We randomly extracted 0.4% of the queries by preserving the original proportion to each type, equivalent to a total of 4,827 queries. This was done for the expedited operation of the experiments, particularly with the restricted access with time interval to Delicious.com. The following is the distribution of the selected queries into different query types: 1,053 single-word queries, 1,909 two-word queries (i.e., each query contains two keywords), 1,325 three-word queries (i.e., each query contains three keywords), and 540 four-word queries (i.e., each query contains four keywords), for a total of 4,827 queries.

Search engine results page. A search engine results pages is the listing of Web pages returned by a search engine in response to a query. In this study, we examined the first ten URLs in each search engine results pages. Each of the 4,827 queries was submitted to the three most widely used search engines, Google, Yahoo!, and Bing. As this study focuses on search engine results only, sponsored links (or paid placement) were excluded from this study, and only non-sponsored search engine results (called organic links) were used for this study. A series of Perl scripts were written to read search engine results pages and to identify, analyse, and extract organic links.

Social tags for organic links. For each organic link identified from search engine results pages, we examined if the linked Web pages that we called "the organic Web pages" were being tagged, and then collected all the social tags assigned to the organic Web pages. Delicious was used as a source of harvesting social tags, as it is known as one of the most popular social tagging sites and has been widely used in the academic research (Gupta *et al.* 2011).

The following describes the data collection process in more details. All the data were collected from May to July 2010.

(1) Examination on whether each organic Web page is tagged. In Delicious, a history of all tagging information such as tags and taggers for a Web page is accessible from a MD5-based converted URL for the target Web page. For example, given a Web page of http://www.delicious.com, all tagging history for the Web page is found at the Delicious site, http://www.delicious.com/url/ea83167936715d3f712f4fb6c78f92d2, where 'ea83167936715d3f712f4fb6c78f92d2' is the MD5-based conversion of the Web page URL (i.e., www.delicious.com) and MD5 is a widely used cryptographic algorithm employed in security applications. The conversion of a URL into an MD5 URL is achieved by using the Digest::MD5 Perl Module and implemented in our Perl scripts. Figure 1 shows a Delicious Web page that records all tagging history for the Delicious homepage. However, if a Web page is not tagged at Delicious, the corresponding Delicious Web page containing the MD5-based converted URL does not exist. Thus, whether an organic link is tagged or not, can be checked by using the MD5 conversion.

(2) Threshold of social tagging. Social tags assigned to the organic Web pages were collected. As shown in Figure 1, an MD5 Delicious Web page displays when the Web page was initially tagged and how many times it has been tagged so far, under the 'History' tab. In this study, a constraint was applied: organic links tagged by at least 100 people are considered for the connection to social tags. In an earlier study of the dynamics of social tags, Golder and Huberman (2006) examined how Web pages were tagged over time and how the frequency distribution of social tags changed. They reported that each tag frequency over the total frequency of all tags used became constant after the first 100 or so tagging activities. This indicates that the frequency ranking of tags is nearly stable after approximately 100 tagging actions. Following the empirical result of the previous study, we also set 100 as the threshold of the minimal number of tagging activities.

(3) Harvest of social tags. The top thirty social tags assigned to Web pages (i.e., resources by organic links) were collected, if the Web pages are tagged by at least 100 people. For the harvest, we accessed MD5-based-converted Delicious Web pages like that shown in Figure 1. A MD5 Delicious Web page contains a full tagging history as well as displays of the top thirty popular social tags used to tag an associated Web page under the 'Top Tags' label placed on the right side of the page. Recently, the Delicious site has been dramatically re-organized and, as a result, the format of the content shown in Figure 1 is no longer available by the time of writing this article. top thirty social tags are compared against queries to examine if a query submitted to search engines for a Web page is also used to tag the same Web page in Delicious. By comparison, we only considered the top popular social tags, because our study focuses on exploring if the query terms are also used as popular social tags (i.e., terms with more consensus across taggers).

The frequency distribution of social tags follows a power-law distribution (Yi and Chan 2009), i.e., an inversed-J shape, showing that fewer tags are more popularly tagged and more tags are less popularly tagged. Thus, we set thirty as a threshold of popularity considering both such a power-law distribution of tags (i.e., we see that top thirty are relatively fewer tags over all the tags being used) and the presence of top thirty popular tags in Delicious.

In summary, a total of 4,827 queries were selected from the Jansen *et al.* study (2005). Each query was submitted to three search engines, i.e., Google, Yahoo!, and Bing. The Webpages from the organic links in search engine results pages were examined to see whether and how they had been tagged in Delicious. Only the Web pages tagged by at least 100 taggers were included in this study. The top thirty popular social tags used for the organic Web pages were harvested. Given two sets of data collected (i.e., user queries and social tags), we analysed the data to investigate our research questions. The following sections report the results of our analysis.

Table 1: Descriptive Statistics

Web query type*	Search engine	a) Number of organic links (Mean/query)	b) Number of organic links tagged in Delicious (Percentage of column a)	c) Number of organic links tagged by at least 100 people in Delicious (Percentage of column b)	d) Number of organic Web pages of which all query terms appear in the top thirty social tags (Percentage of column c)
One-word queries (1,053)	Google	10,102 (9.6%)	5,424 (53.7%)	1,157 (21.3%)	992 (85.7%)
	Yahoo!	9,938 (9.4%)	4,436 (44.6%)	803 (18.1%)	659 (82.1%)
	Bing	12,223 (11.6%)	5,655 (55.3%)	972 (17.2%)	813 (83.6%)
	Average	10,754 (10.2%)	5,172 (48.1%)	977 (18.9%)	821 (84.0%)
Two-word queries	Google	18,379 (9.6%)	8,016 (43.6%)	1,234 (15.4%)	826 (66.9%)
	Yahoo!	18,703 (9.8%)	6,122 (32.7%)	813 (13.3%)	557 (68.5%)
	Bing	20,044 (10.5%)	7,117 (35.5%)	1,004 (14.1%)	690 (68.7%)

(1,909; 2,893)	Average	19,042 (10.0%)	7,085 (37.2%)	1,017 (14.4%)	691 (67.9%)
Three-word queries (1,325; 2,606)	Google	12,459 (9.4%)	4,781 (38.4%)	650 (13.6%)	202 (31.1%)
	Yahoo!	12,014 (9.1%)	3,217 (26.8%)	390 (12.1%)	132 (33.8%)
	Bing	13,934 (10.5%)	4,327 (31.1%)	514 (11.9%)	181 (35.2%)
	Average	12,802 (9.7%)	4,108 (32.1%)	518 (12.6%)	172 (33.2%)
Four-word queries (540; 1,446)	Google	5,031 (9.3%)	1,740 (34.6%)	219 (12.6%)	22 (10.0%)
	Yahoo!	4,856 (9.0%)	1,183 (24.4%)	95 (8.0%)	19 (20.0%)
	Bing	5,576 (10.3%)	1,650 (29.6%)	161 (9.8%)	24 (14.9%)
	Average	5,154 (9.5%)	1,524 (29.6%)	158 (10.4%)	22 (13.9%)
Total queries (4,827)	Google	45,971 (9.5%)	19,961 (43.4%)	3,260 (16.3%)	2,042 (62.6%)
	Yahoo!	45,511 (9.4%)	14,958 (32.9%)	2,101 (14.0%)	1,367 (65.1%)
	Bing	51,777 (10.7%)	18,749 (36.2%)	2,651 (14.1%)	1,708 (64.4%)
	Average	47,752 (9.9%)	17,889 (37.5%)	2,670 (14.9%)	1,706 (63.9%)

* The numbers in column 1 represent the number of queries of the type and the number of unique query terms

Results

Cross coverage of Web resources in Delicious.com and search engine results pages

Table 1 presents the descriptive statistics for the coverage across three different search engines. The number of organic links is shown at the column a of Table 1. It reads with the first row of single-word queries as follows: each of 1,053 single-word queries was submitted to the Google search engine, and a total of 10,102 organic links were collectively obtained for the entire queries. The average number of organic links per query was 9.6. Furthermore, the average number of organic links per query with single-word queries is 10.2. The data in the subsequent rows of the same column also indicated that the average number of organic links per query slightly decreases as more words were used in a single query, i.e., 10.2 for one-word queries, 10.0 for two-word queries, 9.7 for three-word queries, and 9.5 for four-word queries. It was also found that there was little difference in the number of organic links per query between Google (i.e., 9.5 organic links per query) and Yahoo! (i.e., 9.4 organic links per query). However, the gap between Bing (i.e., 10.7 organic links per query submitted to Bing) and any of the other two (i.e., 9.5 or 9.4) is wider than the one between Google and Yahoo!

Column b of Table 1 refers to the Delicious' coverage of the organic Web pages. A chi-squared test on the average frequency shows that the coverage of the Delicious significantly decreases as the number of words in a query increases ($\chi^2 = 1872.50$, df = 3, $p < 0.01$). More specifically, the organic Web pages in search engine results pages by single-word queries (48.1%) are most frequently tagged in Delicious, while those by four-word queries (29.6%) are least frequently tagged. The last row section of the column displays the cross coverage among search engines. A chi-squared test shows that three search engines produce a significantly different coverage ($\chi^2 = 3124.62$, df = 2, $p < 0.01$). More specifically, about 43% of the organic Web pages from Google are found at Delicious, whereas about 33% and 36% of them from Yahoo! and Bing are tagged at Delicious, respectively.

Column c of Table 1 shows the number of organic Web pages that are tagged by at least 100 people in Delicious. The cross coverage pattern for the column c resembles the results found in the column b: (1) the number of the organic Web pages tagged by more than 100 people in Delicious significantly decreases as the number of words in a query increase ($\chi^2 = 371.10$, df = 3, $p < 0.01$), and (2) three search engines are significantly different in terms of the percentage of the organic Web pages tagged by more than 100 people ($\chi^2 = 299.63$, df = 2, $p < 0.01$). About 16% of the organic Web pages in Google are tagged by more than 100 people in Delicious, while around 14% of the organic Web pages in Yahoo! and Bing are tagged by more than 100 people in Delicious.

Associations between Web queries and social tags: analysis of the queries matched with social tags

We further examined how often users' queries appear in top thirty social tags that have been annotated by at least 100 participants. Column d of Table 1 shows the numbers of queries found at the top thirty social tags and the percentages over the number of organic links tagged by at least 100 taggers in Delicious. Note that a total of 977 Web pages tagged in Delicious by at least 100 people (see the one-word queries 'average' row under column c in Table 1). Out of 977, 156 organic Web pages (i.e., 977 - 821 = 156) were not found in the top thirty social tags. A chi-squared test indicates that the average frequencies for the different query types are significantly different ($\chi^2 = 561.10$, df = 3, $p < 0.01$). More specifically, as the number of words in a query increase, the

average percentage of all the query terms that appear in top thirty social tags significantly decreases. Eighty-four percent of the single-word queries appear in top thirty social tags, but only 13.9 percent of the four-word queries are found in top thirty social tags. A decreasing trend is somewhat expected, because all the words in multi-word queries (i.e., two- to four-word queries) should appear in top thirty tags to be included in the analysis. The last rows of Table 1 display the overall results for each search engine. A chi-squared test compared search engines' performance in terms of the frequency of the query terms appearing at top thirty social tags, and revealed that three search engines were not different in the average frequency ($\chi^2 = 3.80$, df = 2, not significant, Google: 62.6%, Bing: 64.4%, and Yahoo!: 65.1%).

Ranking association between Web query terms and social tags: analysis of the queries in a ranked list of social tags

The results with one-, two-, three-, and four-word queries are illustrated in Figures 2-5, respectively. In the case of one-word queries, more than 40% of queries are top-ranked in social tag lists for all the three search engines (see the 1st rank in Figure 2). As more words are used in a query, the percentages of top-ranked queries are gradually decreased (i.e. 20.8% for two-word queries, 15% for three-word queries, and 12.5% for four-word queries (see the 1st rank in Figures 3-5)). For multi-word queries (see Figure 3, 4, and 5), a total number of plotted points are equivalent to the number of queries multiplied by the number of words in a query.

The general rank-frequency distributions for the four different query types are similar in that the distribution curves begin with a significant downward slope, and tend to be more or less flat after passing a threshold where about 80% of queries were accumulated. When the graphs are redrawn by accumulating the frequency percentages, the points that reach the 80% of the accumulated frequency are intercepted by one-word queries at the sixth rank, at the tenth rank for two-word queries, at the fourteenth rank for three-word queries, and at the eleventh rank for four-word queries. Separating from the slope, the data also show that 90% of the queries are accumulated at the twelfth rank for one-word queries, fifteenth rank for two-word queries, twentieth rank for three-word queries, and sixteenth rank for four-word queries.

Additionally, the frequency that Web queries appeared at each rank (i.e., first to thirtieth) of the social tag list was compared for the purpose of examining the performance of the search engines. A series of chi-squared tests indicated that the three search engines produced equivalent performance in general (Google vs. others: $\chi^2 = 21.15$, df = 29, p = 0.85; Yahoo vs. others: $\chi^2 = 15.02$, df = 29, p = 0.98; Bing vs. others: $\chi^2 = 11.56$, df = 29, p = 0.99).

Some best-fit trend lines for Google data are drawn to evaluate the plotted data in Figures 2-5. The equations of the trend lines and R-squared values are also displayed at the top right corner of the figures. As seen from the equations, power functions were used for the trend lines in Figure 2, 3, and 4, but a logarithm function was used for the trend line in Figure 5. The R-squared values ranged from 0.75 to 0.87, indicating the trend lines explained the variance of the data well.

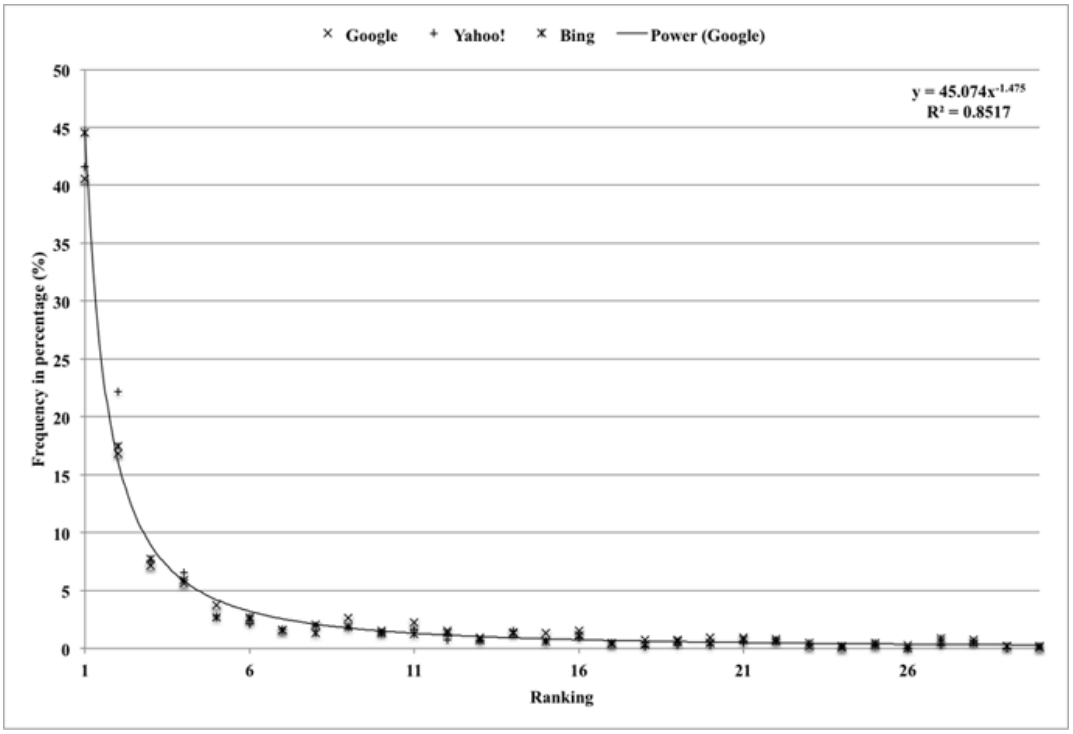

Figure 2: Rank-frequency distribution of one-word queries co-occurring at a list of top thirty social tags with at least 100 taggers

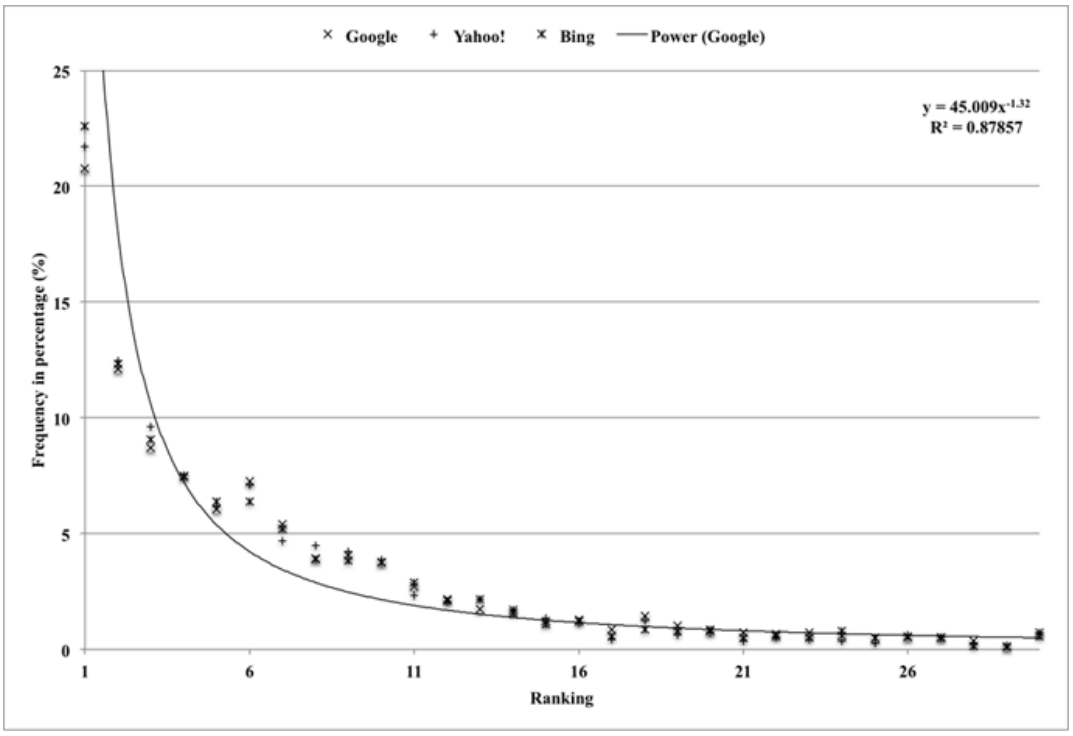

Figure 3: Rank-frequency distribution of two-word queries co-occurring at a list of top thirty social tags with at least 100 taggers

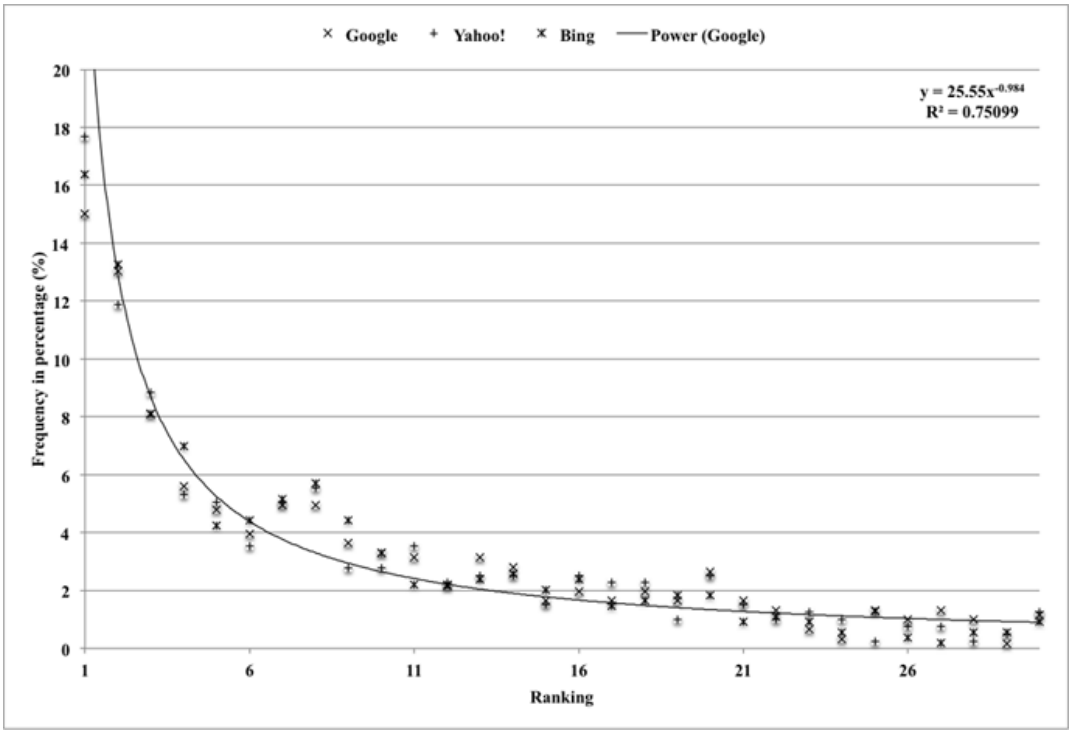

Figure 4: Rank-frequency distribution of three-word queries co-occurring at a list of top thirty social tags with at least 100 taggers

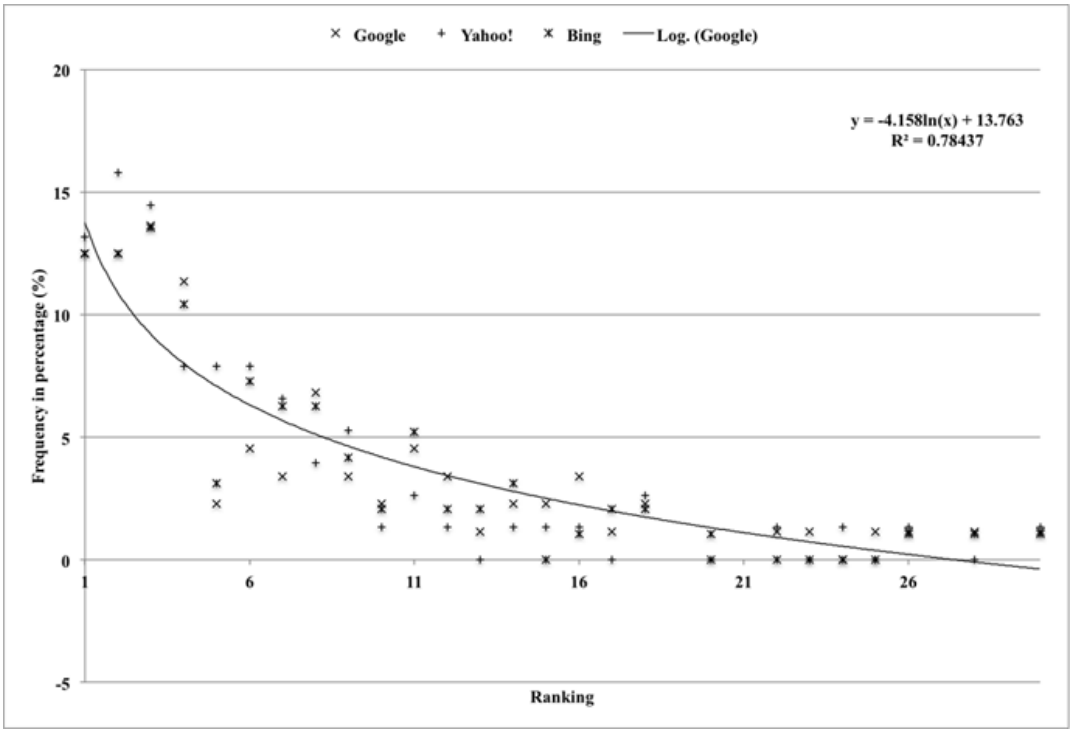

Figure 5: Rank-frequency distribution of four-word queries co-occurring at a list of top thirty social tags with at least 100 taggers

Table 2: Analysis of non-matched query terms

Categories	Percentage (Frequency) One-word query (190)	Two-word query (341)	Three-word query (310)	Four-word query (133)	χ^2Value (df=3)	Total (974)
	χ^2(df=10)=156.35**	χ^2(df=10)=692.88**	χ^2(df=10)=610.33**	χ^2(df=10)=224.99**		χ^2(df=10)=1410.83**
Plural	11.6% (22)	36.7% (125)	38.7% (120)	50.4% (67)	60.99**	34.3% (334)
Proper noun	13.2% (25)	24.9% (85)	30.7% (95)	28.6% (38)	20.42**	25.0% (243)
Misspelling	20.5% (39)	8.8% (30)	7.4% (23)	21.8% (29)	34.67**	12.4% (121)
Abbreviation	14.7% (28)	7.3% (25)	7.1% (22)	17.3% (23)	18.89**	10.1% (98)
Preposition	0.0% (0)	0.3% (1)	5.2% (16)	19.6% (26)	95.10**	4.4% (43)
Number	1.6% (3)	2.1% (7)	4.2% (13)	9.8% (13)	11.31**	3.7% (36)
Ungrammatical composite word	10.5% (20)	0.9% (3)	0.7% (2)	0.8% (1)	60.09**	2.7% (26)
Non-English	3.2% (6)	1.5% (5)	2.6% (8)	2.3% (3)	5.50	2.3% (22)
Article	0.0% (0)	1.2% (4)	3.6% (11)	5.3% (7)	13.99**	2.3% (22)
Conjunction	0.0% (0)	0.0% (0)	1.9% (6)	7.5% (10)	37.46**	1.6% (16)
Slang	1.6% (3)	1.5% (5)	0.3% (1)	3.8% (5)	7.81*	1.4% (14)

*: p<=0.05; **: p <=0.01

Discrepancy analysis: analysis of the non-matching Web queries

In the previous section, we have described how Web queries are matched with social tags in terms of ranking in social tag lists. However, some query terms did not appear in a list of social tags. To explain why some queries did not appear in a list of social tags, we analysed non-matching Web query terms from the functional and structural perspectives. It turned out that 974

(approximately 20%) query terms did not match with social tags derived from search engines (see the Categories row of Table 2: 190 one-word queries, 341 two-word queries, 310 three-word queries, and 133 four-word queries).

A content analytic approach was employed to examine the characteristics of non-matching queries. A total of eleven coding categories was developed, after an iterative process of observing the queries, developing coding categories, and pre-testing some sample queries. The final coding categories used in the analysis are as follows:

- *Abbreviation* for any abbreviation or acronym,
- *Articles* for articles,
- *Conjunction* for conjunctions,
- *Misspelling* for misspelled words and typos,
- *Non-English* for non-English words including Spanish, French, German, Arabic, etc.,
- *Number* for Roman or Arabic numerals,
- *Plural* for plural words,
- *Preposition* for any prepositions,
- *Proper noun* for proper nouns, and
- *Ungrammatical composite word* for any ungrammatically composited words such as "familyfun."
- *Slang*

Two judges were trained for the analysis. To assess the intercoder reliability, two independent judges coded about ten percent of randomly chosen non-matching queries (i.e., a total of 103 query samples; 18 one-word queries, 36 two-word queries, 34 three-word queries, and 15 four-word queries). The inter-coder reliability was assessed by Cohen's kappa, because each judge reached coding decisions in a qualitative and nominal (yes = 1 vs. no = 0) manner. Based on the criteria suggested by Banerjee *et al.* (1999), excellent agreement beyond chance was obtained between the two judges (k = .916). Upon achieving an acceptable reliability coefficient, two judges content-analysed the rest of non-matching queries.

Table 2 shows the descriptive statistics and the results of chi-squared tests. The average frequency of each category appears to be significantly different ($\chi^2 = 1410.83$, df = 11, $p < 0.01$). Among the eleven categories, Plural turned out to be the most frequent with 34.3%, followed by Proper noun (25.0%), Misspelling (12.4%), Abbreviation (10.1%), Preposition (4.4%), Number (3.7%), Ungrammatical composite word (2.7%), Non-English (2.3%), Article (2.3%), Conjunction (1.6%), and Slang (1.4%).

Overall, the result shows that non-matching queries include diverse functional or structural characteristics. However, an interesting observation would be that the contribution of each category varies across the query types. More specifically, as shown in a series of chi-squared tests, Ungrammatical composite word, Misspelling, and Abbreviation more frequently occurs among one-word queries than the other query types, whereas Ungrammatical composite word is the least frequent category in common across the three types other than one-word queries.

Discussion

The objectives of this study were to compare Web queries (i.e., query terms submitted to the three search engines by information users or searchers for the purpose of finding Web resources) and social tags (i.e., terms assigned to the same Web resources by information providers or taggers for the purpose of recognizing, identifying, or sharing Web resources) and to discover how they are associated. Thus, we suggested the four research questions and attempted to answer them.

Research question 1: cross coverage of Web resources between Delicious.com and search engine results pages.

We report that about 38% of the first ten pages in search engine results pages appear at Delicious. In a previous study, Heyman *et al.* (2008) reported 19% coverage of Delicious for the top ten results of Yahoo! search. The two studies, however, differ in the following two areas, which might explain the discrepancy in percentage. First, there is about four-year lag time between the Delicious datasets being used in the two studies, For the four year, Delicious keeps growing and becomes richer. Consequently, we claim that our data reflect tagging activities by a considerably greater number of people. Secondly, in our study, the entire Delicious Website was used to check the coverage on the fly, but, in Heyman's study, the coverage check was made based on the data collected at a specific month consisting of about seven million Delicious URLs. We believe that our on-the-fly matching can produce more accurate result because the most up-to-date, complete Delicious data were fully used.

WorldWideWebSize.com reported that Web pages available from Bing or Yahoo! ranged between ten and twenty billion pages and those from Google ranged between fifteen and forty-five billion pages. Meanwhile, Delicious contains approximately 100 million unique URLs as of September 2007 (Arrington 2007) and the size becomes almost doubled in fourteen months after that

to November 2008, leading to approximately 400 million today. Thus, the relative ratio of the size of Delicious over the size of the searchable Web can be calculated by around 0.01%- (i.e., 400 millions / 40 billions). Our study demonstrated that 38% of the URLs at the first pages of search engine results pages were also covered at Delicious, which is a relatively large number when considering the small ratio between Delicious and Web. Heyman *et al.*'s study showed 9% coverage that was obtained with the top 100 search results, which is different from our result, based on the top ten search results, The two different results may imply that Web pages saved in Delicious are likely to be highly ranked in result pages.

Research question 2: overlap between Web queries and social tags.

We report that at least 60% of search engines' query terms overlap with Delicious tags (refer to the last row of Table 1). Heyman *et al.* (2008), in the same vein, investigated the overlap of popular Delicious tags with AOL (American Online) query terms, but they simply compared the two datasets. Instead, our study made comparisons based on same target URLs that were associated with both social tags and Web queries. In addition, our study was conducted on the fly under real Web searching settings. As a result, data may be more current, realistic and accurate.

Research question 3: ranking association between Web query terms and social tags.

We found that query terms tended to overlap with higher ranked social tags than lower ones, indicating that higher ranked social tags are more likely to be used as query terms for the same Web resources. More specifically, the probability to find the query terms in the top fifteen social tags will be greater than 80%. The finding is of great importance because it offers an empirical threshold in tag ranking within which query terms are likely to be found. In addition, the co-occurring pattern of query terms over the social ranking resembles a power law distribution: higher percentages in co-occurrence at higher ranking and lower percentages at lower ranking. Such a pattern can be a valuable discovery that contributes to expanding the use of social tags to information retrieval applications for Web searches.

Furthermore, we demonstrated a strong association between Web search query terms and social tags, when they co-occur: a higher ranked social tag for a Web resource is more likely to be used as a whole or a part of Web query submitted to search engine for the retrieval of the same resource. The most beneficial application of this finding may be found in search engine optimization, seeking to advance Web pages in rank by increasing their visibility in search engine results pages. Optimizing a Web page involves editing and updating its content, HTML and associated coding to both increase its relevance to specific query keywords and to lower barriers to the indexing activities of search engines. Thus, online marketers or webmasters are recommended to review what keywords are frequently assigned to the same Web page in social tagging systems and include those frequently used keywords in content and metadata. Updating content so as to keep search engines revisiting for crawling back and re-indexing can give additional weight to a Web page. Therefore, adding socially high-ranked keywords to a Web page's metadata, including the title tag and meta description, will improve the relevancy of the Web page, and thus, ultimately, increase traffic to the Webpage.

The primary focus of this study is to discover the overlap and association between social tags and Web queries. The result is not much beyond our initial expectation that there might be a certain level of association. This study contributes to confirm the expectation, with considerable overlaps in percentage and power law-like overlap in ranking. The empirical data for the association can be interpreted as a shared ground between information seeking behavior by Web searchers and information sharing/indexing behavior by social taggers. As Web searching is one of the most popular Internet activities, the Web searcher group must be larger than the social tagger group. Also, the two groups have clearly different goals and purposes: locating information in Web searching versus a combination of sharing, retrieving, summarizing, etc. information in social tagging. Nevertheless, the result implies a shared view on same Web resources between the two different information groups.

Research question 4: discrepancy analysis.

The discrepancy analysis found that both query terms and social tags suffered from similar problems derived from a non-controlled vocabulary. From the functional and structural perspective, non-matched query terms were classified into a number of categories in this study. The results suggest that it may be worthwhile to adopt the following techniques for more overlapping between query terms and social tags: techniques to convert plural cases into singular cases or vice versa, to check and correct misspelled words, and to eliminate some functional words. Similar categories were also identified as problematic areas inherent to social tags (Heckner *et al.* 2007; Kipp and Grant 2006; Golder and Huberman 2006; Li , *et al.* 2008, Yi and Chan 2009). Stemming is a technique to convert inflected forms and sometimes derived forms of a word into a common base form (call stem of the word), and Porter's stemming algorithm, a ruled-based technique, (Porters 1980) is widely used by the information retrieval community. Stop-words lists contain some extremely common words that would appear to be of little value and Salton's 571 list (Ide and Salton, 1971) has been commonly used in text processing (see Manning *et al.* 2008) for more techniques for text processing). This study was conducted in a real setting using real queries submitted by online users to the major search engines and social tags in Delicious. Thus, we claim that the findings about the relationships between social tags and Web queries could

be realistic and directly applicable to other tasks such as search engine optimization. A limitation of this study was that user-clicked information in search engine results pages was not considered and thus it was not an experimental variable in this study. Instead, top ten URLs in results pages are considered in this study. We also would like to point out the time gap between the almost ten-years-old Web queries and the up-to-date tagging vocabulary in Delicious, used in this study. Any new topics or terms require time in order to be established in the Delicious tagging vocabulary. Thus, the gap between the two would not affect any negative effect toward the result of this study. If a recent Web query dataset was employed in this study, it may affect it because of the newly created or used terms that may be involved in the dataset such as new-coined or new-fashioned terms, may not appear in the tagging vocabulary. Thus, such a time gap, although not intended, might be appropriate for the study.

Conclusion and future research

Social tagging data and search engine queries are two separate datasets created for distinct purposes. Based on the postulation, the socially created data have a great potential in assisting or complementing Web queries - we investigated social tags and Web queries. Particularly, we focused on how Web queries and social tags can be linked to, associated with, or complementary to each other. The empirical results can be summarized as follows: first, about 38% of search engine results pages returned from the three major search engines (i.e., Bing, Google, Yahoo!) are also tagged on the social tagging site, Delicious. Secondly, the overlapped percentage between query terms and top thirty social tags that derived from at least 100 taggers is 63.9%. A higher percentage is achieved in queries of fewer words. Thirdly, we also discovered that when query terms co-occur at a social tags lists, as much as 80% of the query terms appear within the top fifteen social tags out of the top thirty social tags, regardless of different query types and search engines. Finally, the co-occurring patterns of query terms and social tags over social ranking are described by a power law.

Social tagging services and systems are still relatively new. The size of socially tagged resources continues to grow, but it is far less than the number of resources indexed by search engines. Our results imply that socially tagged resources are highly ranked in search engine results pages. Social tagging data have been used in information retrieval tasks relative to Web search and such empirical data have indicated the potential to enhance Web search. Along the same lines, this study extends previous studies and contributes to understanding how Web query terms are related to, and associated with, social tags.

In recent Web information retrieval research, some scholars attempted to use social tagging data for improving the quality of Web retrieval performance or Web page ranking models based on clicked Webpages and Wikipedia (Bao *et al.* 2007; Chen and Zhang 2009; Yanbe *et al.* 2007; Schenkel *et al.* 2008). Unlike previous studies, this study attempts to focus on examining the association between social data and Web queries. However as the social data became commonly available, little scientific and empirical research has examined the potential association between social tags and Web queries. This study demonstrates that there is a strong relationship between social tags and Web queries. The finding of this study can be indirectly applicable to the future study of various potential applications including Web query prediction, suggestion, and search engine optimization. Given that the continued growth of social collaboration services and applications, social data can hold more potential in the organization, navigation, retrieval, and manipulation of Web resources. In the near future, rich sets of social tagging data are expected to be more extensively used in information retrieval research and search-engine-related applications. Thus, future studies should investigate the effects of social tagging data on Web search for clicked versus non-clicked search results.

Acknowledgements

The authors would like to thank the anonymous reviewers of this article for their valuable and insightful input and comments.

About the authors

Kwan Yi is an Assistant Professor in the Department of Curriculum and Instruction, Eastern Kentucky University, Kentucky, USA. He received his Ph.D. degree in Information Studies and Master of Science in Computer Science from McGill University, Montreal in Canada. He also obtained his Master of Science in Applied Mathematics from the University of Illinois at Urbana-Champaign, USA. Areas of his research interests are automatic organization and classification of digital information and information retrieval. He can be contacted at kwan.yi@eku.edu

Chan Yun Yoo is an associate professor of Integrated Strategic Communication in the School of Journalism and Telecommunications, University of Kentucky. He holds Ph.D in advertising from the University of Texas at Austin. His areas of research interest include online media advertising and consumer behavior. He can be contacted at cyoo2@email.uky.edu

References

- Alpert, J. & Hajaj, N. (2008). We knew the web was big... Retrieved 5 March, 2012 from http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
- Arrington, M. (2007, September 6). Exclusive: screen shots and feature overview of Delicious 2.0 preview. *TechCrunch*. Retrieved 5 March, 2012 from http://www.techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/ (Archived by WebCite® at http://www.webcitation.org/6ABjdqD0i)
- Baeza-Yates, R. & Tiberi, A. (2007). Extracting semantic relations from query logs. In P. Berkhin, R. Caruana, X. Wu & S. Gaffney, (Eds.), *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 76-85). New York, NY: ACM Press.
- Banerjee, M., Capozzoli, M., McSweeney, L. & Sinha, D. (1999). Beyond kappa: a review of inter-rater agreement measures. *Canadian Journal of Statistics*, **27**(1), 3-23.
- Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B. & Su, Z. (2007). Optimizing web search using social annotations. In C. Williamson, M.E. Zurko, P. Patel-Schneider & P. Shenoy, (Eds.), *Proceedings of the 16th International Conference on the World Wide Web*, (pp. 501-510). New York, NY: ACM Press.
- Bassett, D. & Kumaran, M. (2008). Libraries and Google co-op. *Journal of Library Administration*, **46**(3-4), 181-189.
- Biancalana, C. & Micarelli, A. (2009). Social tagging in query expansion: a new way for personalized Web search. *Proceedings of the IEEE International Conference on Computational Science and Engineering*, (pp. 1060-1065). Los Alamitos, CA: IEEE Computer Society.
- Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, **36**(2), 3-10.
- Carman, N. (2009). *LibraryThing tags and Library of Congress Subject Headings: a comparison of science fiction and fantasy works* Unpublished Master's thesis, Victoria University of Wellington, Wellington, New Zealand. Retrieved 5 March, 2012 from http://researcharchive.vuw.ac.nz/bitstream/handle/10063/1272/thesis.pdf?sequence=1 (Archived by WebCite® at http://www.webcitation.org/6ABvFBaXu)
- Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. & Loreto, V. (2009). Collective dynamics of social annotation. In *Proceedings of the National Academy of Sciences*, **106**(26), 10511-10515. Retrieved 5 March 2012 from http://www.pnas.org/content/106/26/10511.full.pdf+html (Archived by WebCite® at http://www.webcitation.org/6ABvlxK3e)
- Chang, Y.-S., He, K.-Y., Yu, S. & Lu, W.-H. (2006). Identifying user goals from Web search results. In L. Carr, D. De Roure, A. Iyengar, C. Goble & M. Dahlin, (Eds.), *Proceedings of the 15th International Conference on the World Wide Web*, (pp. 1038-1041). New York, NY: ACM Press.
- Chen, S.-Y. & Zhang, Yi. (2009). Improve Web search ranking with social tagging. In Francisco Manuel Carrero, José María Gómez, Borja Monsalve, Enrique Puertas & José Carlos Cortizo (Eds.). *Proceedings of the 1st International Workshop on Mining Social Media, 13th Conference of the Spanish Association for Artificial Intelligence*, Madrid: Bubok Publishing S.L.
- Collins-Thompson, K. & Callan, J. (2005). Query expansion using random walk models. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates & N. Ziviani, (Eds.), *Proceedings of the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 704-711). New York, NY: ACM Press.
- Craswell, N., Hawking, D. & Robertson, S. (2001). Effective site finding using link anchor information. In D.H. Kraft, W.B. Croft, D.J. Harper & J. Zobel, (Eds.), *Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 250-257). New York, NY: ACM Press.
- Cui, H., Wen, J.R., Nie, J.Y. & Ma, W.Y. (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering*, **15**(4), 829-839.
- Ding, Y., Jacob, E., Fried, M., Toma, I., Yan, E., Foo, S. & Milojevic, S. (2010). Upper Tag Ontology (UTO) for integrating social tagging data. *Journal of the American Society for Information Science and Technology*, **61**(3), 505-521.
- Drott, M. C. (2002). Indexing aids at corporate websites: the use of robots.txt and META tags *Information Processing & Management*, **38**(2), 209-219.
- eBizMBA. (2011, October). Top 15 most popular Web 2.0 websites. Retrieved 5 March, 2012 from http://www.ebizmba.com/articles/web-2.0-websites (Archived by WebCite® at http://www.webcitation.org/6ABw6NwBw)
- Eiron, N. & McCurley, K.S. (2003). Analysis of anchor text for Web search. In C. Clarke, G. Cormack, J. Callan, D. Hawking & A. Smeaton, (Eds.), *Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 459-460). New York, NY: ACM Press.
- Golder, S. & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, **32**(2), 198-208.
- Gupta, M., Li, R., Yin, Z. & Han, J. (2011). An overview of social tagging and applications. In C.C. Aggarwal, (Ed.), *Social network data analytics* (pp. 447-497). New York, NY: Springer.

- Hammond, T., Hannay, T., Lund, B. & Scott, J. (2005). Social bookmarking tools (I): A general review. *D-Lib Magazine*, **11**(4). Retrieved 5 March 2012 from http://www.dlib.org/dlib/april05/hammond/04hammond.html (Archived by WebCite® at http://www.webcitation.org/6ABwyuuf8)
- Heckner, M., Muhlbacher, S. & Wolff, C. (2007). Tagging tagging: a classification model for user keywords in scientific bibliography management systems. *Journal of Digital Information*, **9**(2). Retrieved 27 August, 2012 from http://journals.tdl.org/jodi/article/view/246/208 (Archived by WebCite® at http://www.webcitation.org/6AEMzPHVP)
- Heymann, P., Koutrika, G. & Garcia-Molina, H. (2008). Can social bookmarking improve web search? In M. Najork, A. Broder & S. Chakrabarti, (Eds.), *Proceedings of the international conference on Web search and web data mining*, (pp. 195-206). New York, NY: ACM Press.
- Hu, J., Wang, G., Lochovsky, F., Sun, J.T. & Chen, Z. (2009). Understanding user's query intent with Wikipedia. In D. Lassner, D.D. Roure & A. Lyengar, (Eds.), *Proceedings of the 18th International Conference on the World Wide Web*, (pp. 471-480). New York, NY: ACM Press.
- Ide, E. & Salton, G. (1971). Interactive search strategies and dynamic file organization. In G. Salton, (Ed.), *The SMART retrieval system experiments in automatic document processing* (pp. 373-393). Englewood Cliffs, NJ: Prentice-Hall Inc.
- Jansen, B. J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, **42**(1) 248-263.
- Jansen, B. J., Spink, A. & Pederson, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, **56**(6), 559-570.
- Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing & Management*, **36**(2) 207-227.
- Kipp, M.E.I. & Grant, C.D. (2006). Patterns and inconsistencies in collaborative tagging systems: an examination of tagging practices. *Proceedings of American Society for Information Science and Technology*, **43**(1), 1-8. Retrieved 5 March 2012 from http://eprints.rclis.org/bitstream/10760/8720/1/KippCampbellASIST.pdf (Archived by WebCite® at http://www.webcitation.org/6ABxMOnMB)
- Lawrence, S. & Giles, C.L. (2000). Accessibility of information on the Web. *Intelligence*, **11**(1), 32-39.
- Lee, U., Liu, Z. & Cho, J. (2005). Automatic identification of user goals in Web search. In A. Ellis, T. Hagino, F. Douglis & P. Raghavan, (Eds.), *Proceedings of the 14th International Conference on the World Wide Web*, (pp. 391-400). New York, NY: ACM Press.
- Li, X., Guo, L. & Zhao, Y.E. (2008). Tag-based social interest discovery. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins & X. Zhang, (Eds.), *Proceedings of the 17th International Conference on the World Wide Web*, (pp. 675-684). New York, NY: ACM Press.
- Manning, C.D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*, New York, NY: Cambridge University Press.
- Mei, Q., Zhou, D. & Church, K. (2008). Query suggestion using hitting time. In J.G. Shanaham, S. Amer-Yahia, I. Manolescu, Y. Zhang, D.A. Evans, A. Kolcz, K.-S. Choi & A. Chowdury, (Eds.), *Proceedings of the 17th ACM conference on Information and Knowledge Management*, (pp. 469-478). New York, NY: ACM Press.
- O'Reilly, T. (2005). What is Web 2.0? - Design patterns and business models for the next generation of software. Retrieved 5 March 2012 from http://oreilly.com/web2/archive/what-is-web-20.html (Archived by WebCite® at http://www.webcitation.org/6ACGaIQhX)
- O'Reilly, T. & Battelle, J. (2009). Web squared: Web 2.0 five years on. Web 2.0 Summit. Retrieved 5 March 2012 from http://assets.en.oreilly.com/1/event/28/web2009_websquared-whitepaper.pdf (Archived by WebCite® at http://www.webcitation.org/6ABxgElSQ)
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web. Stanford, CA: Stanford University, InfoLab. Retrieved 25 August 2012 from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=5668173FAA589B65E379868395A9A769?doi=10.1.1.31.1768&rep=rep1&type=pdf (Archived by WebCite® at http://www.webcitation.org/6ABy6GutJ)
- Pew Internet Project. (2010, December 16). *Generations 2010.* Washington, DC: Pew Research Center. Retrieved 5 March 2012 from: http://www.pewinternet.org/~/media//Files/Reports/2010/PIP_Generations_and_Tech10.pdf (Archived by WebCite® at http://www.webcitation.org/6ACGp6NKp)
- Porter, M.F. (1980). An algorithm for suffix stripping, *Program*, **14**(3), 130-137.
- Rose, D. & Levinson, D. (2004). Understanding user goals in web search. In S. Feldman, M. Uretsky, M. Najork & C. Wills, (Eds.), *Proceedings of the 13th International Conference on the World Wide Web*, (pp. 13-19). New York, NY: ACM Press.
- Schwarz, B. (2006, May 11). Google Co-op: What is it? *Search Engine Roundtable*. Retrieved 5 March, 2012 from http://www.seroundtable.com/archives/003796.html (Archived by WebCite® at http://www.webcitation.org/6AByPV38Q)
- Seaver, B. (2007, March 5). Web 2.0 stats - Fascinating growth in blogs, video, MySpace. Retrieved 5 March, 2012 from http://microexplosion.blogspot.com/2007/03/web-20-stats-fascinating-growth-in.html (Archived by WebCite® at http://www.webcitation.org/6AByXn7ac)

- Song, Y, Zhou, D. & He, L.-W. (2011). Post-ranking query suggestion by diversifying search results. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua & W.B. Croft, (Eds.), *Proceedings of the 34th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 815-824). New York, NY: ACM Press.
- Spiteri, L.F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*, **26**(3), 13-24.
- Wu, L.F. (2011). The accelerating growth of online tagging systems. *European Physical Journal B - Condensed Matter and Complex Systems*, **83**(2), 283-287. Retrieved 5 March ,2012 from http://www.epj.org/_pdf/HP_EPJB_accelerating_growth.pdf (Archived by WebCite® at http://www.webcitation.org/6AByex0bg)
- Xu, Y., Jones, J.F. & Wang, B. (2009). Query dependent pseudo-relevance feedback based on Wikipedia. In J. Allan, J. Aslam, M. Sanderson, C.-X. Zhai & J. Zobel, (Eds.), *Proceedings of the 32nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, (pp. 59-66). New York, NY: ACM Press.
- Yanbe, Y., Jatowt, A., Nakamura, S. & Tanaka, K. (2007). Can social bookmarking enhance search in the web? In E. Rasmussen, R.R. Larson, E. Toms, & S. Sugimoto, (Eds.), *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries*, (pp. 107-116). New York, NY: ACM Press.
- Yi, K. (2009). A study of evaluating the value of social tags as indexing terms. In S. Chu, W. Ritter, & S. Hawamdeh (Eds.), *Managing Knowledge for Global and Collaborative Innovations* (pp. 221-232), Hackensack, NJ: World Scientific Publishing.
- Yi, K. & Chan, L. M. (2009). Linking folksonomy to Library of Congress Subject Headings: an exploratory study. *Journal of Documentation*, **65**(6), 872-900.

How to cite this paper

Yi, K. & Yoo, C.Y. (2012). "An empirical examination of the associations between social tags and Web queries" *Information Research*, 17(3) paper 527. [Available at http://InformationR.net/ir/17-3/paper527.html]

- Contents |
- Author index |
- Subject index |
- Search |
- Home